GWAS to Sequencing:
Sequencing Think Tank and Training Workshop
for the Genes, Environment and Health Initiative (GEI)

January 18-19, 2007

Attachments

Appendix 1: Meeting agenda
Appendix 2: List of questions posed to workshop participants
Appendix 3: Meeting roster

The Genes, Environment and Health Initiative (GEI; see http://www.gei.nih.gov/) seeks a better understanding of how genetic factors, environmental exposures, and gene-environment interactions contribute to human disease. A primary goal of the genetic component of GEI is to develop methods for comprehensive analysis of genetic variation in regions identified by well-conducted genome-wide association studies in order to determine the best candidate genes and variants for functional studies. This report summarizes the discussions held at a workshop organized by the trans-NIH GEI Genetics Sub-committee to address the questions "When and how should sequencing be used to follow up on whole-genome association studies?"

Specific technical, strategic, and analytical questions were posed to the participants (see Appendix 2) to guide the discussion:

- What criteria should be used for picking a genomic region to sequence in depth?
- What determines the boundaries of a region? Within a region, what should be the coverage and depth of sequence analysis?
- How many individuals should be sequenced? How should they be chosen?
- How can the sequencing be accomplished most efficiently?
- How will newer sequencing technologies alter the current paradigms?
- What defines an "interesting" variant?
- What are the most important statistical or analytic issues for analyzing the data?

Reasons for sequencing after genome-wide association studies

The workshop participants agreed that there are at least three distinct (but related) approaches for using genomic sequencing to follow up on a genome-wide association study (GWAS). The three approaches are complementary and could represent succeeding stages of analysis. The choice of approaches for a study will depend on the specific goals of the study, the amount of additional knowledge of the regions involved, and the available resources (both financial and samples). The three approaches are:

1. Use of sequencing to discover the "functional element" that is responsible for the phenotype of interest in the region of an association hit. Only a small proportion of the variants in the causative gene will actually be causal, and only a fraction of those will be readily identifiable as variants that result in amino acid changes that obviously would affect the function of the gene product (i.e., truncate it or otherwise disrupt the coding sequence). However, the easiest way to try to identify

the functional elements of interest is to find such signature “smoking gun” variants. Samples from many individuals usually need to be sequenced, but only in exons, and the analysis is straightforward. Although the identification of putative “smoking guns” provides strong clues to the identity of the causal functional elements, follow-up experiments are required for confirmation.

The genes identified this way are a good place to start to narrow down a broad association peak, but it must be recognized that this approach likely will give false negatives (because some causal genetic elements will not have exonic variants) and false positives (because some “smoking guns” will turn out not to have functional effects). Since we do not know what proportion of causal functional elements can be detected in this manner, establishing this proportion would be valuable for designing future studies, but this will require large-scale sequencing as outlined in approach 2.

Approach 1 will only identify the functional elements that should be examined more comprehensively. It will not find all variants or all functional variants in a gene, particularly those that are not in coding sequences. The strong LD across many genomic regions often extends past exon boundaries; thus sequencing approach 2 will be necessary to characterize the patterns of variation to allow the prioritization of genetic elements and variants that can then be subjected to the functional analyses of approach 3.

2. Use of sequencing to determine the complete genotype/phenotype correlation structure in the “functional element”. The overall aim is to deeply characterize, within a population, most of the variants in an identified region associated with a phenotype. Note that the region might have been identified or narrowed as outlined in approach 1, or there may be other reasons to be confident that the region contains the functional element of interest, such as excellent linkage or association data that narrow the region to one or only a very few genes, or the existence of an obvious candidate gene. If the data show that a few adjacent genes have LD or are all associated with a phenotype, so that the actual functional element is narrowed down but not individually identified, it may still be worth sequencing through the region to find the genotype/phenotype correlation structure. A judgment to proceed could depend on the significance of the disease, the cost of producing the data, and other factors.

   This approach will entail looking at all the sequence in these regions, not just exons. It was suggested in discussion that the boundaries could be as much as 0.5 Mb on both sides of each region, perhaps even more in some cases, depending on the LD in the region. This extensive sequencing would allow detection of variants that are in upstream or downstream regulatory regions and introns as well as in exons. Although this approach could be limited to looking only for variants that might contribute to the original association hit (such as by looking for all variants in LD with the tag SNPs that identified the hit), more generally the goal will be to fully describe the genetic variation, both common and rare, in the regions associated with the phenotypes of interest.

   Pragmatic limits on the size of the sample that can be analyzed will determine the degree to which rare variants and those that contribute only modest effects can be found. Variants identified by this approach will still require testing to determine whether they are causal or merely associated. However, a full understanding of the pattern of variation within the region based on sequence data from many individuals can provide insight into how to choose variants for subsequent functional testing (see approach 3).

3. Use of sequencing to provide clues that will guide functional studies undertaken to understand the mechanism and identify which genetic elements and variants are actually causal. This is a goal that the sequencing data facilitate, but does not involve a different type of sequencing than approach 2. Approach 2 finds as much information as possible from the sequence data, which will include most common variants and many rare ones. This information includes the frequency distribution of variants, the LD patterns among them, and the set of haplotypes showing the co-transmitted variants across the exons, promoter regions, and other regulatory regions. Once sets of variants are identified containing putative causal variants that cannot be individually teased out because of LD, experimental functional studies will be needed to identify the functional genetic elements and variants and elucidate the mechanisms of how they contribute to the phenotype.

General recommendations

1. No matter what the reason for sequencing, the evidence for an association must be strong, generally based on replication studies following the GWAS, before beginning to sequence. The workshop participants cited many cases where there was controversy about a claim of association; sequencing resources are best used in cases with good evidence for proceeding. Consideration should include the quality of the initial study and replication studies, and the strength of the signal.

2. It should not be assumed that a variant discovered in a region of association is a causal variant. This point is perhaps obvious, but enough mistakes have been made in the past that it was worth emphasizing.

3. Regions of association may be chosen jointly for follow-up, based on the results from multiple studies.

Specific considerations for each approach

Approach 1: To find the functional elements

Most GEI studies will start with this approach.

1. The sample size has to be large enough to give the investigator a reasonable chance of finding a rare "smoking gun" variant. It should be emphasized that in this case we are not looking for all the variants. Rather, the goal will be to identify variants that alter the protein structure (pre-terminal stop or amino acid substitution/ deletion) in a way that allows the conclusion that the coding sequence thus identified is a strong candidate for affecting the phenotype of interest. Since such variants are expected to be rare, this will generally entail examining hundreds or thousands of samples. The search could be staged, for example by beginning with a pilot study in which a few hundred samples were sequenced; if the pilot study were to show no obvious functional exonic variants, more samples could then be added. There was strong agreement that the initial studies should err on the side of sequencing more samples. To demonstrate the utility of this approach, the GEI program should start with a few studies that will allow sequencing of samples from many individuals (1000 cases and 1000 controls). Doing so will also generate data sets large enough to enable development of much-needed new analytical approaches and tools for sequence analysis and genotype/phenotype association studies.
2. Samples should be chosen from the extremes of the phenotype distribution. Choosing complementary phenotypes, e.g., late vs. early onset, as well as extremes of exposures, should also be considered.

3. The utility of selecting samples based on haplotypes is currently a matter of debate. GEI could include studies with samples chosen by both strategies (selected by haplotype or not selected by haplotype) to address this question. The HapMap data are useful for selecting which haplotypes to use. Sequence data will help define haplotype structure more finely, but the issue of phasing is formidable and will require better analytical tools. A more complete reference sequence, which would include haplotypes, could also facilitate these studies.
4. The point at which the study will stop (i.e., no new samples will be added) must be defined. This is a cost-benefit analysis to avoid overspending in cases where a “smoking gun” variant is not found initially.
5. The target regions to be sequenced should be approximately 1 Mb. However, a larger or smaller region may be justified based on information such as LD in the region.
6. Better algorithms for recognizing interesting variants are needed. The automated mutation detection sensitivity for single-base variants may not be sufficient, especially in non-coding regions. Hence, large data sets will be useful for developing new tools and approaches.
7. Use of allele-specific primers can lead to haplotypes being missed.
8. The new sequencing technologies (454, Solexa) may be particularly appropriate for this approach, particularly if pools of samples are sequenced, which would increase the chances of finding rare variants.
9. A subset of these studies should involve sequencing more extensively than just in the exons, at least in a few hundred individuals, to gain insight into what might be missed with smaller sample sizes and only exonic sequencing.

Approach 2: To understand the complete genotype/phenotype correlation structure in the functional elements

When the region containing the functional element has been defined with some confidence (by sequencing approach 1 or by other means), then comprehensive sequencing in many individuals will provide information on most of the variants in the region (common and rare, SNP as well as structural), the LD patterns among them, and how the pattern of variation relates to genetic elements.

Analyses of these sequence data may provide clues to infer function, such as phenotype associations with common variants, excess frequencies of rare variants, or patterns of variation indicating natural selection.

1. The sample size should be on the order of thousands to find and characterize enough rare variants that could contribute to disease in populations.
2. The target interval should be ~1 Mb, but could be larger (as discussed above).
3. Looking at samples from multiple ethnic groups may show differences in LD that, in some circumstances, could be useful for fine mapping.
4. Some of these studies should be done soon, perhaps independently of GEI. The NHGRI-supported allelic spectrum project does aim to address some of these issues. In addition, there are some other possible gene regions of interest for these early studies, including AMD, IL23R, CAPON, IBD5, and perhaps another 20 that are known at this time.
5. When sequence variants are found in a region, follow-up genotyping can provide adequate power in a cost-effective manner for the analysis of genotype-phenotype correlations.

6. This approach will provide data that are a good initial challenge for tools designed to identify and characterize rare variants. Development of such tools will be essential for analysis of whole genome data when they become available in the future.

Approach 3: To provide clues to guide functional studies to find the functional elements and variants and elucidate the mechanisms

Analysis of large amounts of sequence data, combined with other information about important functional regions (e.g., from ENCODE), may provide clues about which functional studies should be done to gain insight into mechanisms. Experimental functional analysis will be needed to follow up the candidate functional genes and variants identified as being associated with the phenotypes of interest, always remembering that there may be many variants in a region with high LD among them. The biological information obtained may include gene regulation, and gene-gene and gene-environment interactions. These functional studies of how particular genes and variants contribute to phenotypes may not scale as high-throughput genomic approaches, although they may use genomic resources.

<u>Other considerations and recommendations</u>

1. Development of methods to recover targeted regions (~1 Mb) in the genome would be a technical breakthrough for the field.
2. In considering how to prioritize among good signals, the significance and robustness of the candidate signal based on GWAS or candidate gene studies with adequately powered replication studies must be taken into account. A major factor should be the putative biological relevance or function of a gene in the region, if known, although the scope of this knowledge is at present quite limited. Additional considerations can include region size, significance of the disorder, potential payoff, and the potential for immediate diagnosis or intervention.
3. The discovery of new variants, especially rare ones, requires very high accuracy in the calling of SNPs and structural variants, and in the sequence assembly.
4. New methods of analysis of sequence data are urgently needed, particularly for the interpretation of the majority of variants in non-coding regions. Large data sets should be generated to drive the development of new analytical tools that will allow prioritization of variants for functional analysis or testing in follow-up studies, and of *in silico* tools that predict functional effects. Methods are needed for finding associations of common variants with phenotypes using the LD structure, detecting associations of phenotypes with rare variants or excess numbers of rare variants, and identifying patterns of selection indicating functional effects of variants or haplotypes.
5. Currently, sequencing the exons of 50 genes in 1000 samples would cost about $1 million, and sequencing 1 Mb in 1000 samples would cost about $4-6 million. A cost-effective, efficient technology for typing a small number of variants (such as would be identified in a sequencing study) in a moderate or large-scale follow-up study is needed. New technologies could be compared in tandem.
6. The data should be made broadly available, through public databases or data enclaves (if required by IRB constraints), with appropriate curation.